Detection of Outliers in Multivariate Data: A Method Based on Influence Eigen

Nazrina Aziz

***Abstract*—Outliers can be defined simply as an observation (or a subset of observations) that is isolated from the other observations in the data set. There are two main reasons that motivate people to find outliers; the first is the researchers intention. The second is the effects of an outlier on analyses. This article does not differentiate between the various justifications for outlier detection. The aim is to advise the analyst of observations that are considerably different from the majority. This article focuses on the identification of outliers using the eigenstructure of S and $\mathbf{S}_{(i)}$ in terms of eigenvalues, eigenvectors and principle component. Note that $\mathbf{S}_{(i)}$ is the sample covariance matrix of data matrix $\mathbf{X}_{(i)}$, where the subscript i in parentheses is read as "with observation i removed from X". The idea of using the eigenstructure as a tool for identification of outliers is motivated by Maximum Eigen Difference (MED). MED is the method for identification of outliers proposed by [1]. This method utilises the maximum eigenvalue and the corresponding eigenvector. It is noted that examination of observations effect on the maximum eigenvalue is very significant. The technique for identification of outliers discuss in this article is applicable to a wide variety of settings. In this article, observations that are located far away from the remaining data are considered to be outliers.**

***Index Terms*—outliers, eigenvalue, eigenvector, covariance matrix, principle component.**

I. Introduction

THIS article focuses on the identification of outliers using the eigenstructure of $\mathbf{S}$ and $\mathbf{S}_{(i)}$ in terms of eigenvalues, eigenvectors and principle component. Note that $\mathbf{S}_{(i)}$ is the sample covariance matrix of data matrix $\mathbf{X}_{(i)}$, where the subscript i in parentheses is read as "with observation i removed from $\mathbf{X}$".

The idea of using the eigenstructure as a tool for identification of outliers is motivated by Maximum Eigen Difference (MED). This method utilizes the maximum eigenvalue and the corresponding eigenvector. It is noted that examination of the observations effect on the maximum eigenvalue is very significant. The reason is that outliers that lie in the direction close to the maximum eigenvalue or vice versa, will change the maximum eigenvalue [1]. The maximum eigenvalue contains maximum variance, therefore, the outliers detected by the maximum eigenvalue have a greater effect on variance, and they need extra attention.

The article is organized as follows: Section II describes a general idea of the influence eigenvalues and eigenvectors. Section III explains the technique, influence eigen for identification of outlier. Three different scenarios are considered to generate the data set from the multivariate distributions in this article are described in Section IV. Finally, Section V provides illustrative examples before presenting the conclusion.

II. Influence eigenvalues and eigenvectors

Some statistical methods are concerned with eigenstructure problems and a few statistics are the functions of eigenvalues in multivariate analysis. A test statistic is considered as a function of eigenvalues of a transition matrix to test a Markov chain for independence [5] and eigenstructure methods are applied to study the co-linear problem in multivariate linear regression [7].

Now, consider the influence of eigenvalues λ_j and eigenvectors v_j for matrix $\mathbf{X}^T\mathbf{X}$ where $\mathbf{X}$ is an $n \times p$ observation matrix consisting of n observations for p variables.

If ith row of matrix $\mathbf{X}$ is deleted, one can write it as $\mathbf{X}_{(i)}$ where the subscript i in parentheses is read as "with observation i is removed from $\mathbf{X}$", i.e. the ith row of $\mathbf{X}$ is x_i^T then $\mathbf{X}_{(i)}^T\mathbf{X}_{(i)} = \mathbf{X}^T\mathbf{X} - x_i x_i^T$. Let $\mathbf{X}^T\mathbf{X}$ have the eigenvalues-eigenvectors pairs

$$(\lambda_1, v_1), (\lambda_2, v_2), ..., (\lambda_p, v_p),$$

and the eigenvalues are in descending order

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p, \tag{1}$$

and let $\mathbf{X}_{(i)}^T\mathbf{X}_{(i)}$ have the eigenvalues and eigenvectors pairs

$$(\lambda_{1(i)}, v_{1(i)}), (\lambda_{2(i)}, v_{2(i)}), ..., (\lambda_{p(i)}, v_{p(i)}),$$

and the eigenvalues are also in descending order

$$\lambda_{1(i)} \geq \lambda_{2(i)} \geq ... \geq \lambda_{p(i)}. \tag{2}$$

Define,

$$\mathbf{V}_{(i)} = [v_{1(i)}, v_{2(i)}, \dots, v_{p(i)}], \tag{3}$$

and

$$\begin{aligned}\mathbf{V}_{(i)}^T\mathbf{S}_{(i)}\mathbf{V}_{(i)} &= \mathbf{V}_{(i)}^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})\mathbf{V}_{(i)} \\ &= \text{diag}[\lambda_{1(i)}, \lambda_{2(i)}, \dots, \lambda_{p(i)}] \\ &= \mathbf{\Lambda_{(i)}}. \end{aligned} \tag{4}$$

Then influence functions of eigenvalues λ_j and eigenvectors v_j are given respectively by [4] as follows:

$$IF(x; \lambda_j) = (x^T v_j)^2 - \lambda_j \tag{5}$$

and

$$IF(x; v_j) = -x^T v_j \sum_{k \neq j} x^T v_k (\lambda_k - \lambda_j)^{-1} v_k. \tag{6}$$

If one wishes to examine the ith observation's influence on the eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$, it is easy to remove the ith observation from the full data set and then compare the eigenvalues and eigenvectors of the remaining data with that of the complete data.

Lemma 1: The properties of eigenvalues and eigenvectors are given as follows:

1) $\lambda_j \geq \lambda_{j(i)}$;

Manuscript received March 7, 2012; revised April 14, 2012. Nazrina Aziz is a senior lecturer in School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia (email: nazrina@uum.edu.my).

2) The relationship of eigenvalues λ_j and $\lambda_{j(i)}$ is given by [1]:

$$\lambda_{j(i)} = \lambda_j - \frac{1}{n-1}(l_{ij}^2 - \lambda_j) - \frac{1}{2(n-1)^2} l_{ij}^2 \Big[1 + \sum_{k \neq j} \frac{l_{ij}^2}{\lambda_k - \lambda_j}\Big] + O(\frac{1}{n^3}), \quad (7)$$

where $l_{ij} = (x_i - \bar{x})^T v_j$;

3) The relationship between eigenvectors of v_j and $v_{j(i)}$ is obtained based on the observation matrix $\mathbf{X}$ given by [1] as follows:

$$\begin{aligned} & v_{j(i)} \\ & = v_j + \frac{l_{ij}}{n-1} \sum_{k \neq j} \frac{l_{ik} v_k}{\lambda_k - \lambda_j} \\ & \quad - \frac{1}{2(n-1)^2} \sum_{k \neq j} \Big[\frac{l_{ij}^2 l_{ik}^2 v_j}{(\lambda_k - \lambda_j)^2} - \frac{2 l_{ik}^2 l_{ij}}{(\lambda_k - \lambda_j)} \sum_{k \neq j} \frac{l_{ik} v_k}{\lambda_k - \lambda_j} \\ & \quad + \frac{2 l_{ij}^3 l_{ik} v_k}{(\lambda_k - \lambda_j)^2} \Big] + O(\frac{1}{n^3}). \qquad (8) \end{aligned}$$

Proof: (i) $\lambda_j \geq \lambda_{j(i)}$ is obtained from the following matrix operations: It is noted that

$$\mathbf{X}^T \mathbf{X} = \mathbf{X}_{(i)}^T \mathbf{X}_{(i)} + x_i x_i^T,$$

where $\mathbf{X}^T\mathbf{X}$, $\mathbf{X}_{(i)}^T \mathbf{X}_{(i)}$ and $x_i x_i^T$ are symmetric matrices and $x_i x_i^T$ is of rank unity, there exists on an orthogonal matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^T (x_i x_i^T) \mathbf{Q} \quad = \quad \begin{pmatrix} s & 0 \\ 0 & 0 \end{pmatrix},$$

where s is the unique non-zero eigenvalues of $x_i x_i^T$, and consider

$$\mathbf{Q}^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)}) \mathbf{Q} \quad = \quad \begin{pmatrix} t & c^T \\ c & \mathbf{X}_*^T \mathbf{X}_* \end{pmatrix},$$

then there is an orthogonal matrix $\mathbf{P}_{(k-1)(k-1)}$ so that

$$\mathbf{P}^T (\mathbf{X}_*^T \mathbf{X}_*) \mathbf{P} = \Lambda_* = diag\{\lambda_1, \lambda_2, ... \lambda_{k-1}\}$$

and one can define an orthogonal matrix

$$\mathbf{G} \quad = \quad \mathbf{Q} \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P} \end{pmatrix},$$

then

$$\begin{aligned} \mathbf{G}^T (\mathbf{X}^T \mathbf{X}) \mathbf{G} \quad & = \quad \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P}^T \end{pmatrix} \mathbf{Q}^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)}) \mathbf{Q} \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P} \end{pmatrix} \\ & + \quad \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P}^T \end{pmatrix} \mathbf{Q}^T (x_i x_i^T) \mathbf{Q} \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P} \end{pmatrix} \\ & = \quad \begin{pmatrix} t + s & c^T \mathbf{P} \\ \mathbf{P}^T c & \Lambda_* \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} \sum_{j=1}^{k} \lambda_j & = t + s + \sum_{i=1}^{k-1} \lambda_i \\ & = t + \sum_{i=1}^{k-1} \lambda_i + s \\ & = \sum_{j=1} \lambda_{j(i)} + s. \qquad (9) \end{aligned}$$

Note that $s \geq o$, and $\lambda_j \geq \lambda_{j(i)}$ is obtained for any $i = 1, 2, ..., n$. ■

III. Influence eigen for identification of outlier

Let the sample covariance matrix be

$$\mathbf{S} = \frac{1}{n} \mathbf{X}^T (\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T) \mathbf{X}, \qquad (10)$$

where $\mathbf{1}$ is the n-vector of ones and $\mathbf{I}_n$ is the identity matrix of $n \times n$. If $\mathbf{X}_{(I)}$ and $\mathbf{S}_{(I)}$ are the data matrix and sample covariance matrix, respectively, when the m observations are deleted and the subscript I in parentheses is read as “with a set of m observations I removed from $\mathbf{X}$”, note that $I = \{i_1, i_2, \ldots, i_m\}$ where $1 \leq i_j \leq n$ and $j = 1, 2, \ldots, m$.

Therefore, one has

$$\mathbf{S}_{(I)} = \frac{1}{n-m} \mathbf{X}_{(I)}^T (\mathbf{I}_{n-m} - \frac{1}{n-m} \mathbf{1}_{n-m} \mathbf{1}_{n-m}^T) \mathbf{X}_{(I)} \quad (11)$$

and

$$\mathbf{S}_I = \frac{1}{m} \mathbf{X}_I^T (\mathbf{I}_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T) \mathbf{X}_I. \qquad (12)$$

Lemma 2: It is noted that

1) The relationship among $\mathbf{S}$, $\mathbf{S}_I$ and $\mathbf{S}_{(I)}$ is given as follows:
$\mathbf{S}_{(I)} = \frac{n}{n-m}\mathbf{S} - \frac{nm}{(n-m)^2}[\frac{n-m}{n}\mathbf{S}_I + (\bar{x}_I - \bar{x})(\bar{x}_I - \bar{x})^T]$;
2) If let $I = \{i\}$ with a single observation, then
$$\mathbf{S}_{(i)} = \tfrac{n}{n-1}\mathbf{S} - \tfrac{n}{(n-1)^2}(x_i - \bar{x})(x_i - \bar{x})^T.$$

Proof: (i) Suppose that equations 10-12 are biased estimates, they can be used to developed unbiased estimates as in lemma 2.

$$\begin{aligned} & (n-m)\mathbf{S}_{(I)} \\ & = X_{(I)}^T \Big(I_{n-m} - \frac{\mathbf{1}}{n-m} \mathbf{1}_{n-m} \mathbf{1}_{n-m}^T \Big) \mathbf{X}_{(I)} \\ & = \mathbf{X}_{(I)}^T \mathbf{X}_{(I)} - \frac{\mathbf{1}}{n-m} \mathbf{X}_{(I)}^T \mathbf{1}_{n-m}^T \mathbf{1}_{n-m} \mathbf{X}_{(I)} \\ & = n\mathbf{S} - \frac{nm}{n-m} (\bar{x} - \bar{x}_I)(\bar{x} - \bar{x}_I)^T + m \bar{x}_I \bar{x}_I^T - \mathbf{X}_I^T \mathbf{X}_I \qquad (13) \end{aligned}$$

Now simplify equation 13 as follows:

$$\begin{aligned} & (n-m)\mathbf{S}_{(I)} \\ & = n\mathbf{S} - \frac{nm}{n-m} (\bar{x} - \bar{x}_I)(\bar{x} - \bar{x}_I)^T - m \Big(\frac{\mathbf{X}_I^T \mathbf{X}_I}{m} - \bar{x}_I \bar{x}_I^T \Big) \\ & = n\mathbf{S} - \frac{nm}{n-m} (\bar{x} - \bar{x}_I)(\bar{x} - \bar{x}_I)^T - m \mathbf{S}_I \qquad (14) \end{aligned}$$

(ii) By using equation 14, one can get the relationship between $\mathbf{S}$, $\mathbf{S}_I$ and $\mathbf{S}_{(I)}$ in (i) where $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ and $\bar{x}_I = \frac{\sum_{i \subset I} x_i}{m}$ represent the mean vector of all observations and the mean vector of the observations indexed by I respectively. Next, replace $m = 1$ in the following equation

$$\begin{aligned} & (n-1)\mathbf{S}_{(I)} \\ & = n\mathbf{S} - \frac{nm}{n-m} (\bar{x} - \bar{x}_I)(\bar{x} - \bar{x}_I)^T - m \Big(\frac{\mathbf{X}_I^T \mathbf{X}_I}{m} - \frac{\mathbf{X}_I^T \mathbf{1}_m \mathbf{1}_m^T \mathbf{X}_I}{m^2} \Big) \end{aligned}$$

hence one can find equation (ii) in lemma 2 as following

$$\begin{aligned} & (n-1)\mathbf{S}_{(i)} \\ & = n\mathbf{S} - \frac{n}{n-1} (\bar{x} - \bar{x}_i)(\bar{x} - \bar{x}_i)^T - 1 \Big(\frac{\mathbf{X}_I^T \mathbf{X}_I}{1} - \frac{\mathbf{X}_I^T \mathbf{X}_I}{1^2} \Big) \\ & = n\mathbf{S} - \frac{n}{n-1} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \end{aligned}$$

This completes the proof of lemma 2. ■

Lemma 3: Let $\{(\lambda_j, v_j), j = 1, 2, ..., p\}$ be the pair of eigenvalues and eigenvectors of sample covariance matrix $\mathbf{S}$. $\{(\lambda_{j(i)}, v_{j(i)}), i = 1, 2, ..., n\}$ be the pair of eigenvalues and eigenvectors of covariance matrix $\mathbf{S}_{(i)}$. One now has

1) $\lambda_{j(i)} = \frac{n}{n-1}\lambda_j - \frac{n}{(n-1)^2}||x_i - \bar{x}_i||^2 G_i$

   where the weights G_i satisfy $0 \leq G_i \leq 1$ and $\sum_i G_i = 1$;

2) $\frac{n}{n-1}\lambda_{j+1} \leq \lambda_{j(i)} \leq \frac{n}{n-1}\lambda_j$, $j = 1, 2, ..., p$.

Proof: It follows immediately from Theorem 1 in [6]

1) Denote $\alpha_i = (x_i - \bar{x})/\|x_i - \bar{x}\|$ and from lemma 2, one has

$$\mathbf{S}_{(i)} = \frac{n}{n-1}\mathbf{S} - \frac{n}{(n-1)^2}(x_i - \bar{x})(x_i - \bar{x})^T. \quad (15)$$

Replace α_i in equation 15 which implies

$$\mathbf{S}_{(i)} = \frac{n}{n-1}\mathbf{S} - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2 \alpha_i \alpha_i^T. \quad (16)$$

Given that

$$\frac{n}{n-1}\lambda_j - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2 \leq \lambda_{j(i)} \leq \frac{n}{n-1}\lambda_j,$$
$$j = 1, 2, \ldots, p.$$

Thus, the weights G_i satisfies $0 \leq G_i \leq 1$ such that

$$\lambda_{j(i)} = \frac{n}{n-1}\lambda_j - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2 G_j. \quad (17)$$

Now, the preceding equation can be written as

$$\text{trace}\, S_{(i)} = \frac{n}{n-1}\text{trace}\, S - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2. \quad (18)$$

From equation 17, one has

$$\begin{aligned}\text{trace}\, S_{(i)} &= \sum_{j=1}^{p} \lambda_{j(i)} \\ &= \frac{n}{n-1}\text{trace}\, S - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2 \sum_{i=1}^{p} G_i. \quad (19)\end{aligned}$$

As a consequence of equations 18 and 19, one has $\sum_{j=1}^{p} G_j = 1$.

2) The proof is given in Corollary 1 and 2 in [6]

■

Theorem 4: The influence eigen j for each observation i can be denoted by

$$\Delta^*_{j(i)} = (x_i^T v_j)^2 + \sum_{\substack{k=1 \\ k \neq i}}^{n} \Big\{(v_j + v_{j(i)})^T x_k x_k^T (v_j + v_{j(i)})\Big\}, \quad (20)$$

where $j = 1, 2, \ldots, p$.

Proof: According to [2] an influence interpretation of the Euclidean distance can be considered as the total of influence eigen:

$$\begin{aligned}&\frac{n}{(n-1)}(x_i - \bar{x})^T(x_i - \bar{x}) \\ &= \sum_{j=1}^{p}\Big\{\frac{1}{n-1}\Big(l_{ij}^2 - \lambda_j\Big) + \frac{1}{2(n-1)^2} l_{ij}^2\Big(1 + \sum_{k \neq j}\frac{l_{ij}^2}{\lambda_k - \lambda_j}\Big)\Big\}.\end{aligned} \quad (21)$$

By using the relationship of influence eigenstructure in lemma 1, equation 21 can be re-written as follows:

$$\begin{aligned}&\frac{n}{(n-1)}(x_i - \bar{x})^T(x_i - \bar{x}) \\ &= \sum_{j=1}^{p}\Big[(x_i^T v_j)^2 + \sum_{\substack{k=1 \\ k \neq i}}^{n}\Big\{(v_j + v_{j(i)})^T x_k x_k^T (v_j + v_{j(i)})\Big\}\Big].\end{aligned} \quad (22)$$

From equation 22, the influence eigen j for each observation i can be denoted by

$$\Delta^*_{j(i)} = (x_i^T v_j)^2 + \sum_{\substack{k=1 \\ k \neq i}}^{n}\Big\{(v_j + v_{j(i)})^T x_k x_k^T (v_j + v_{j(i)})\Big\}, \quad (23)$$

where $j = 1, 2, \ldots, p$.

■

However, if one considers the influence eigen j on I, thus Theorem 4 now becomes

$$\begin{aligned}\Delta^*_{j(I)} = \frac{-m}{n-m}\sum_{k=1}^{n}(x_k^T v_j)^2 - \frac{nm}{(n-m)^2} v_j^T \times \\ \Big[\frac{n-m}{n}\mathbf{S}_I + (\bar{x}_I - \bar{x})(\bar{x}_I - \bar{x})^T\Big] v_j\end{aligned} \quad (24)$$

Suppose that the influence of an observation, i.e. an outlier on statistics such as jth eigenvalues, λ_j or eigenvectors, v_j of a sample covariance matrix is simply the change in λ_j or v_j when the ith observation is deleted from the sample.

Recall that this article considers the maximum eigenvalue and the corresponding eigenvector as the object of interest. From equation 1, it is given that

$$\begin{aligned}max\{\lambda_1, \lambda_2, \ldots, \lambda_p\} &= \lambda_{max} \\ &= \lambda_1,\end{aligned} \quad (25)$$

where λ_1 corresponds to v_1. Now, let $j = 1$, and equation 20 becomes

$$\Delta^*_{1(i)} = (x_i^T v_1)^2 + \sum_{\substack{k=1 \\ k \neq i}}^{n}\Big\{(v_1 + v_{1(i)})^T x_k x_k^T (v_1 + v_{1(i)})\Big\}. \quad (26)$$

Therefore, one can consider the influence eigen, $\Delta^*_{1(i)}$ as a tool to identify a potential influence observation, i.e. outlier in data matrix $\mathbf{X}$. Perhaps the best advice is that the observation that is obviously more extreme than most of the remaining observations in the data set should be examined.

As a consequence, by using $\Delta^*_{1(i)}$, potential outliers in $\mathbf{X}$ can be identified by plotting the index plot of $\{i, \Delta^*_{1(i)}\}$. Note that ith observation can be considered as a potential outlier if it is located further away than the remaining observations in the data set. By using lemma 2, 3 and equation 26 the algorithm for influence eigen, $\Delta^*_{1(i)}$ is given as follows:

- Step 1 : Generate the sample covariance matrix $\mathbf{S}$ and $\mathbf{S}_{(i)}$;

- Step 2 : Compute the eigenstructure of $\mathbf{S}$ and $\mathbf{S}_{(i)}$. Denote the eigenstructure of $\mathbf{S}$ and $\mathbf{S}_{(i)}$ as $\{\boldsymbol{\Lambda}, \mathbf{V}\}$ and $\{\boldsymbol{\Lambda}_{\mathbf{(i)}}, \mathbf{V}_{\mathbf{(i)}}\}$ respectively.
- Step 3 : Choose the maximum eigenvalue and the corresponding eigenvector pair, $max\{\lambda_j, v_j\}$ and $max\{\lambda_{j(i)}, v_{j(i)}\}$ of $\{\boldsymbol{\Lambda}, \mathbf{V}\}$ and $\{\boldsymbol{\Lambda}_{\mathbf{(i)}}, \mathbf{V}_{\mathbf{(i)}}\}$ respectively, i.e. $\{\lambda_1, v_1\}$ and $\{\lambda_{1(i)}, v_{1(i)}\}$;
- Step 4 : Compute $\Delta^*_{1(i)} = (x_i^T v_1)^2 + \sum_{\substack{k=1 \\ k \neq i}}^{n} \Big\{(v_1 + v_{1(i)})^T x_k x_k^T (v_1 + v_{1(i)})\Big\}$ for each observation;
- Step 5 : Develop the index plot of $\{i, \Delta^*_{1(i)}\}$, $i = 1, 2, \ldots, n$.

The outliers that are detectable from the index plot are those which inflate variance and covariance. If an outlier is the cause of a large increase in variances of the original variables, then it must be extreme on those variables [2]. Thus, one can identify it by looking at the index plot.

IV. Simulation data set

The influence eigen is tested on the simulation data set.Three different scenarios are considered to generate the data set from the multivariate distributions with sample size of 3005 and dimensions of 100. The sample size contains 5 outliers.

A. Scenario 1: outliers with the same shapes but different locations

There are 2 conditions considered in the first scenario:

- Condition 1 : A random vector of $x_1, x_2, \ldots, x_n$ is drawn from a $p-$variate normal distribution with mean vector μ and positive definite covariance matrix $\mathbf{\Sigma}$, i.e. $N(\mu, \mathbf{\Sigma})$. Next $x_1^*, x_2^*, \ldots, x_m^*$ is another random sample drawn from a $p-$variate normal distribution with mean vector $\mu_{\mathbf{c1}}$ and a similar covariance matrix $\mathbf{\Sigma}$, i.e. $N(\mu_{\mathbf{c1}}, \mathbf{\Sigma})$. Note that m is the number of outliers. Later these two sets of data vector are merged;
- Condition 2 : The $x_1, x_2, \ldots, x_n$ random vector is developed as in condition 1. However, $x_1^*, x_2^*, \ldots, x_m^*$ is constructed by using $N(\mu_{\mathbf{c2}}, \mathbf{\Sigma})$, which is closer to the majority of data parental distribution in condition 1, i.e. $\mu_{\mathbf{c2}} < \mu_{\mathbf{c1}}$;

B. Scenario 2: outliers with different shapes and different locations

In scenario 2, $x_1, x_2, \ldots, x_n$ is a random vector drawn for $p-$ variate normal distribution with mean vector μ and positive definite matrix $\mathbf{\Sigma}$ and $x_1^*, x_2^*, \ldots, x_m^*$ is another set of random vector from $p-$ variate distribution with mean vector $\mu_{\mathbf{s2}}$ and covariance matrix $\mathbf{\Sigma}_{\mathbf{s2}}$. Note that $\mu \neq \mu_{\mathbf{s2}}$ and $\mathbf{\Sigma} \neq \mathbf{\Sigma}_{\mathbf{s2}}$.

C. Scenario 3: outlier from a different probability law

Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from $p-$ variate normal distribution with mean vector μ and positive definite covariance matrix $\mathbf{\Sigma}$. Now generate $x_1^*, x_2^*, \ldots, x_m^*$ drawn from $p-$ variate student t distribution with z degrees of freedom and correlation matrix $\mathbf{\Sigma}_{\mathbf{s3}}$. Note that $\mathbf{\Sigma} \neq \mathbf{\Sigma}_{\mathbf{s3}}$.

V. Illustration by simulation data set

The technique is used on the simulate data set which generated following three scenarios described in previous section. Recall that the last 5 observations in the simulated data set are the outliers. Fig. 1 clearly displays these 5 outliers at the top of each index plot of $\Delta^*_{1(i)}$. It is noted that in the index plot, there is a very large gap between the outliers and the remaining observations, i.e. good data. Generation of a multivariate data set from a population where the good and bad data are closer to each other probably causes the suggested technique not to perform. Nonetheless, Fig. 2 indicates all 5 observations that are supposed to be outliers in the data set are considered for condition 2.

Recall that scenario 2 generated a data set with different shapes and different locations. It is known the last 5 observations in this data set are the outliers. Fig. 3 clearly displays these 5 outliers at the top of each index plot of $\Delta^*_{1(i)}$. It is noted there is a large gap between the outliers and the remaining observations.

Fig. 4 also shows that $\Delta^*_{1(i)}$ is capable of identifying outliers in a high-dimensional data set that contains outliers coming from different probability of laws. Note that outliers are denoted within the black circles.

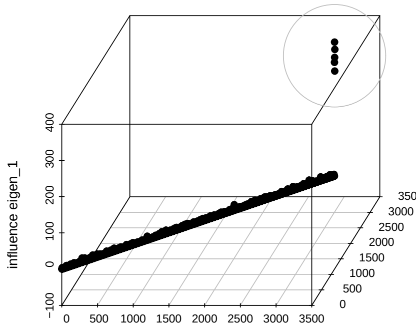

Fig. 1. 3D Scatterplot for Condition 1, Scenario 1.

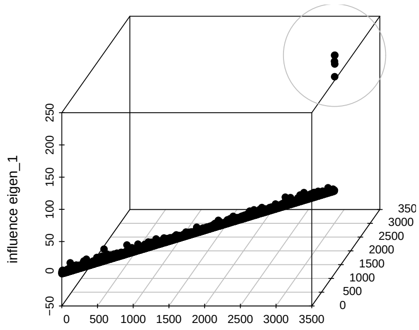

Fig. 2. 3D Scatterplot for Condition 2, Scenario 1.

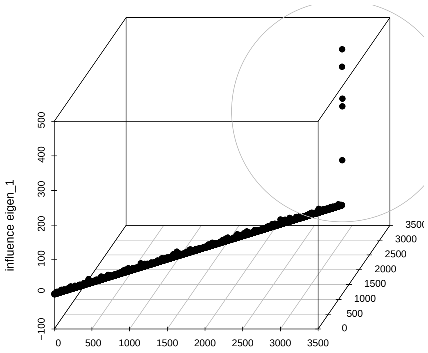

Fig. 3. 3D Scatterplot for Scenario 2.

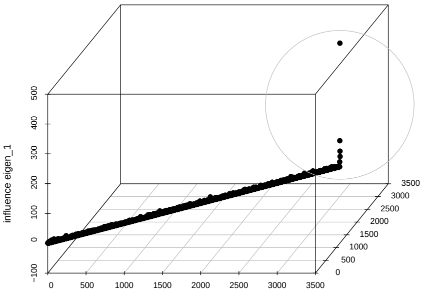

Fig. 4. 3D Scatterplot for Scenario 3.

VI. Conclusion

Sometimes, the identification of outliers is the main objective of the analysis, and whether to remove the outliers or for them to be down-weighted prior to fitting a non-robust model. This article does not differentiate between the various justifications for outlier detection. The aim is to advise the analyst of observations that are considerably different from the majority. Note that the technique in this article is, therefore, exploratory. The technique used in this article is performed on large data set. In this article, observations that are far away from the remaining data are considered to be outliers.

If the ith observation is a potential outlier, their values for $\Delta^*_{1(i)}$ are all situated at the top of the index plot; see illustration of index plots in previous section. This is because an outlier causes $\lambda_1 - \lambda_{1(i)}$ values to be larger than other observations. Note that $\lambda_{1(i)}$ value is smaller for an outlier. This follows that $\Delta^*_{1(i)}$ become larger.

References

[1] Gao. S, Li. G, Wang. D. Q, ”A new approach for detecting multivariate outliers,” *Communication in Statistics-Theory and Method* vol.34, 2005, pp. 1857–1865.

[2] Gnanadesikan. R, Kettenring. J. R, ”Robust estimates, residuals, and outlier detection with multiresponse data,” *Biometrics* vol.28, 1972, pp. 81–124.

[3] Jolliffe. I. T, *Principal Component Analysis*. Springer-Verlag, New York, 1972.

[4] Radhakrishnan. R, Kshirsagar. A. M, ”Influence functions for certain parameters in multivariate analysis,” *Communication in Statistics-Theory and Method*, vol. 10, 1981, pp. 515–529.

[5] Wang. D. Q, Scott. D. J, ”Testing a markov chain for independence,” *Communication in Statistics-Theory and Method*, vol. 18, 1989, pp. 4085–4103.

[6] Wang. S.G, Liski. E. P, ”Effects of observations on the eigensystem of a sample covariance matrix,” *Journal of Statistical Planning and Inference*, vol. 36, 1993, pp. 215–226.

[7] Wang. S. G, Nyquist. H, ”Effects on the eigenstructure of a data matrix when deleting an observation,” C*omputational Statistics and Data Analysis*, vol. 11, 1991, pp. 179–188.